Optimization of Carbó Molecular Similarity Index Using Gradient Methods

ALAN J. McMAHON and PAUL M. KING*

Department of Chemistry, Birkbeck College, University of London, Gordon House, 29 Gordon Square, London WC1H 0PP, United Kingdom

Received 30 November 1995; accepted 26 April 1996

ABSTRACT

A steepest descent method for optimizing the Carbó molecular similarity index was implemented and evaluated. Comparisons were made between this procedure and the extensively used simplex method. Several data sets were considered, and in each case the gradient method showed a substantial improvement in the time taken for the optimization to converge while comparable similarity values were obtained. In some cases, performance enhancements of up to an order of magnitude were observed. © 1997 by John Wiley & Sons, Inc.

Introduction

Molecular similarity methods provide a simple quantitative measure of how similar two molecules are. These methods have found widespread use within the general area of rational drug design.[1] In the modeling of unknown receptor sites molecular similarity is used to superimpose molecules known to bind to a particular site and to elicit specific responses, so that the shape and charge distribution of the site can be deduced from its complimentarity to the aligned molecules.[2] The superimposition of molecules based upon their molecular similarity also finds application in Quantitative Structure Activity Relationship (QSAR) studies, where attempts are made to correlate observed chemical or biological response with certain molecular properties and characteristics of the molecules.[3–6] The design and molecular modeling of bioisosteres and transition-state mimics also relies upon the quantitative nature of molecular similarity to validate the results of such studies. Molecular similarity is also used as a scoring function in the screening of data bases for potential drug molecules having related behavior to a known ''lead'' compound. The related concept of molecular ''dissimilarity'' is used to quantify how different the enantiomeric forms of a chiral compound are. This can produce a chirality coefficient that is often found to correlate well with the eudismic ratio of enantiomeric pairs. There are several recent reviews of molecular similarity to which the interested reader is referred.[7–9]

* Author to whom correspondence should be addressed. E-mail: p.king@chem.bbk.ac.uk

CCC 0192-8651 / 97 / 020151-08

All of the applications cited above depend crucially on the superimposition of the two molecules upon which a similarity calculation is being performed. Molecular similarity indices are often seen to be very sensitive to the alignment of the molecules concerned. The simplest means of superimposition involves aligning the molecules, using, for example, centers of mass, centers of charge, moments of inertia, electrostatic multipole moments, least-squares fitting of selected atomic positions, etc. However, there is no guarantee that superimposition based on these relatively simple measures of geometric shape or charge distribution will generate alignment of molecules in order to produce optimal, or even realistic, similarity values. Given one molecule, A, and another, B, with position coordinates $\mathbf{R}_A$ and $\mathbf{R}_B$, respectively, the similarity is a function of both the molecular conformation of each molecule and their relative position and orientation.

$$\text{Similarity of } A \text{ and } B = S_{AB} = S(\mathbf{R}_A, \mathbf{R}_B, \mathbf{R}_{\text{COM}}, \boldsymbol{\Theta}), \quad (1)$$

where $\mathbf{R}_{\text{COM}}$ is the displacement vector of the centers of mass and $\boldsymbol{\Theta}$ is the vector of three Euler angles defining the relative orientation of the molecules. The function S_{AB} will thus map out a multidimensional surface containing many stationary points and, ideally, a single maximum that represents the maximum similarity value corresponding, in a similarity sense, to optimal alignment of the two molecules. The best way, therefore, to align molecules such that they produce an optimal similarity value is to search the similarity space spanned by the conformations and relative configurations of the two molecules. The simplest search procedures assume that the two molecules are rigid, perhaps in independently energy minimized structures, so that the search only entails exploring the 6-dimensional space of relative displacement and orientation. Earliest applications of molecular similarity adopted this approach.[10,11] Ideally, however, the ability of molecules to change conformation should be included in the optimization process, and flexible-fitting methods were employed.[12] These methods essentially add the internal degrees of freedom of each molecule to the search variables, and include a molecular mechanics penalty function to validate optimization moves.

To date, optimization of molecular similarity has relied upon algorithms that use solely function evaluations, primarily within the simplex approach.[13] Methods aimed at improving the speed of similarity calculations, as outlined below, provided the opportunity of using searching procedures that utilize gradient information. In this article we present the first application of searching using gradient-based methods.

Background

A quantitative measure of the molecular similarity between two molecules, A and B, was first defined by Carbó et al.[14] as follows:

$$R_{AB} = \frac{\int \rho_A(\mathbf{r})\,\rho_B(\mathbf{r})\,d\mathbf{r}}{\left(\int \rho_A(\mathbf{r})^2\,d\mathbf{r}\right)^{1/2}\left(\int \rho_B(\mathbf{r})^2\,d\mathbf{r}\right)^{1/2}}, \quad (2)$$

where $\rho_A(\mathbf{r})$ and $\rho_B(\mathbf{r})$ are the electron densities of molecules A and B, respectively. However, this is not a unique functional form for the index and neither is the choice of electron density the only, or indeed optimal, choice for the molecular property upon which to base the similarity measure. An alternative and commonly used index was introduced by Hodgkin and Richards[10] that is more sensitive to the magnitudes of the electron densities than the Carbó index.

Most practical calculations of molecular similarity on large molecules and molecules of potential pharmacological importance have not used the electron density as the molecular property of interest, but rather the electrostatic potential or the molecular shape. This is because many applications on large data bases of molecules and/or large molecules require a classical rather than quantum mechanical description of the electrostatics of the molecules in question. The classical electrostatic potential of a point charge distribution provides a description of how the molecule will be recognized by others at medium- to long-range distances. In the discussion that follows, we demonstrate the differentiation of the Carbó electrostatic potential similarity index, and the use of the derivatives obtained to optimize the similarity. Our software also allows the differentiation of the Hodgkin index and the use of shape, rather than electrostatic potential for either of these (Carbó or Hodgkin) indices.

In the present work we consider the calculation of the Carbó similarity index using the electrostatic potential of the molecules instead of the electron

densities. This involves replacing $\rho(\mathbf{r})$ of eq. (2) by $V(\mathbf{r})$. The electrostatic potential has the advantage that it is straightforward to calculate classically using atom-centered point charges:

$$V(\mathbf{r}) = \sum_{i=1}^{N} \frac{q_i}{|\mathbf{r} - \mathbf{R}_i|}, \tag{3}$$

where q_i is the charge on atom i centered at position $\mathbf{R}_i$ and N is the number of atoms in the molecule. Evaluation of the integrals required for the evaluation of molecular similarity was originally performed numerically on a grid. However, the use of a 2- or 3-Gaussian expansion for the $1/r$ terms of eq. (3) allows the grid-based determination of the electrostatic potential to be replaced by analytic evaluations that make the calculation 2 orders of magnitude faster. Using a Gaussian expansion the electrostatic potential becomes[15]

$$V(\mathbf{r}) = \sum_{i=1}^{N} q_i \sum_{j=1}^{N_{\text{Gauss}}} \gamma_j e^{-\alpha_j(\mathbf{r} - \mathbf{R}_i)^2} \tag{4}$$

where N_{Gauss} is the number of Gaussians used in the expansion. All the integrals required for the evaluation of the molecular similarity can now be determined analytically using standard results.[16]

The optimization of the molecular similarity, i.e., obtaining the optimal superimposition of two molecules, involves the use of an algorithm to search the space formed by the relative orientation of both molecules for the best similarity value. When numerical integration of the relevant integrals is utilized only optimization procedures based upon function evaluations or that involve the numerical calculation of derivatives can be used. Traditionally the simplex algorithm has found widespread use. However, if use is made of the Gaussian expansion in the electrostatic potential [eq. (4)], not only can the similarity index be calculated analytically but so too can the derivatives of the similarity index with respect to the relative orientation of the two molecules. It was this realization upon which the work described in this article is primarily based. The availability of rapidly calculated analytic derivatives enables a whole range of gradient-based search procedures to be utilized.[17] In this study we used the simplest of the gradient-based methods, that of steepest descents.

If molecule B is considered to be stationary while molecule A is moving, the expression for the derivative of the Carbó index can be written as follows:

$$\left(\frac{\partial R_{AB}}{\partial \mathbf{R}_i^A}\right) = \left(\frac{R_{AB}}{I_{AB}}\right)\left(\frac{\partial I_{AB}}{\partial \mathbf{R}_i^A}\right), \tag{5}$$

where

$$I_{AB} = \sum_{i=1}^{N_A}\sum_{j=1}^{N_B} q_i^A q_j^B \left(\sum_{k=1}^{N_t} s_k e^{(t_k|\mathbf{R}_i^A - \mathbf{R}_j^B|^2)}\right), \tag{6}$$

and

$$\left(\frac{\partial I_{AB}}{\partial \mathbf{R}_i^A}\right) = \sum_{j=1}^{N_B} q_i^A q_j^B \sum_{k=1}^{N_t} 2 s_k t_k \left(\mathbf{R}_i^A - \mathbf{R}_j^B\right) e^{(t_k|\mathbf{R}_i^A - \mathbf{R}_j^B|^2)}. \tag{7}$$

Here $N_t = 3$ for a 2-Gaussian expansion and $N_t = 6$ for a 3-Gaussian expansion. The various constants in the expansion of eq. (6) are given in Table I. It is straightforward to transform the atom-based derivative of the similarity index expressed in eq. (5) to forces acting upon the center of mass and torques about a molecule-centered axis frame. These derivatives can then be used in a searching algorithm to optimize the superimposition of the two molecules. As outlined earlier, we can also use this method for optimization of the Hodgkin index and for the optimization of shape similarity.[18]

Computer Implementation

All molecules were built using the PIMMS[19] molecular modeling package. The structures were subsequently optimized by semiempirical quantum mechanical calculations using the MNDO[20] Hamiltonian, while atom-centered partial charges were obtained by fitting to the electrostatic potential calculated at this level of theory according to the method of Besler et al.[21] These calculations were performed using the MOPAC[22] program.

The program for molecular similarity calculations was written in ANSI standard C on a Silicon Graphics Indigo-2 workstation. Prior to optimization of the similarity index the centers of mass of the lead and the analogue under consideration were superimposed. The optimization of the similarity index proceeds by minimizing the negative of the similarity and hence maximizing the similarity itself. This was performed using a basic steepest descent algorithm. This is clearly not a sophisticated algorithm but serves as an initial

TABLE I.
Terms in Gaussian Expansion Evaluation of Molecular Similarity.

	2-Gaussian Expansion		3-Gaussian Expansion	
k	s_k	t_k	s_k	t_k
1	$\gamma_1^2\left(\frac{\pi}{2\alpha_1}\right)^{3/2}$	$-\left(\frac{\alpha_1}{2}\right)$	$\gamma_1^2\left(\frac{\pi}{2\alpha_1}\right)^{3/2}$	$-\left(\frac{\alpha_1}{2}\right)$
2	$2\gamma_1\gamma_2\left(\frac{\pi}{\alpha_1+\alpha_2}\right)^{3/2}$	$-\left(\frac{\alpha_1\alpha_2}{\alpha_1+\alpha_2}\right)$	$2\gamma_1\gamma_2\left(\frac{\pi}{\alpha_1+\alpha_2}\right)^{3/2}$	$-\left(\frac{\alpha_1\alpha_2}{\alpha_1+\alpha_2}\right)$
3	$\gamma_2^2\left(\frac{\pi}{2\alpha_2}\right)^{3/2}$	$-\left(\frac{\alpha_2}{2}\right)$	$2\gamma_1\gamma_3\left(\frac{\pi}{\alpha_1+\alpha_3}\right)^{3/2}$	$-\left(\frac{\alpha_1\alpha_3}{\alpha_1+\alpha_3}\right)$
4			$\gamma_2^2\left(\frac{\pi}{2\alpha_2}\right)^{3/2}$	$-\left(\frac{\alpha_2}{2}\right)$
5			$2\gamma_2\gamma_3\left(\frac{\pi}{\alpha_2+\alpha_3}\right)^{3/2}$	$-\left(\frac{\alpha_2\alpha_3}{\alpha_2+\alpha_3}\right)$
6			$\gamma_3^2\left(\frac{\pi}{2\alpha_3}\right)^{3/2}$	$-\left(\frac{\alpha_3}{2}\right)$

benchmark for a gradient-based method with which to compare previously used simplex methods, and also subsequent more powerful gradient-based searching methods. Optimizations were performed in the 6-dimensional space spanned by the displacement of the centers of mass and the three angles specifying relative orientations. Both molecules were considered to be rigid throughout the optimization process. The steepest descent method is implemented by updating the position of the moving molecule according to the formula

$$\mathbf{R}_{\text{new}} = \mathbf{R}_{\text{old}} + \alpha\mathbf{g}, \tag{8}$$

where $\mathbf{g}$ is the normalized rigid-body force acting on the moving molecule and α is a variable that forms the step size in the search procedure. The optimization ends when either the similarity value is greater than 0.99, all partial derivatives are within some value (conv_tol, typically 10^{-4}) of zero, or the maximum number of iterations are exceeded, i.e.,

FOR n = 1 TO MAX_ITERATIONS
BEGIN
  calculate similarity
  calculate gradients
  IF (similarity has improved)

$$\alpha = \alpha * 1.2$$

  ELSE

$$\alpha = \alpha/2.0$$

  IF (similarity > 0.99 OR function converged)

    BREAK

  alter position of moving_molecule [eq. (8)]
END

The performance of the gradient-based optimizations and the similarity values obtained was compared against the frequently used simplex method. This was performed on two sets of data previously described in similarity works,[6,15] and one new set.[23] The molecules used in this study are shown in Figures 1–3.

Results

Table II gives the calculated similarity values for the three series of molecules considered in this study. The simplex average column represents averaged results from 100 simplex calculations started with different random seeds. The simplex best column lists the highest single similarity value obtained from the 100 different runs. The steepest descent column gives the results from a single optimization using the methods described here.

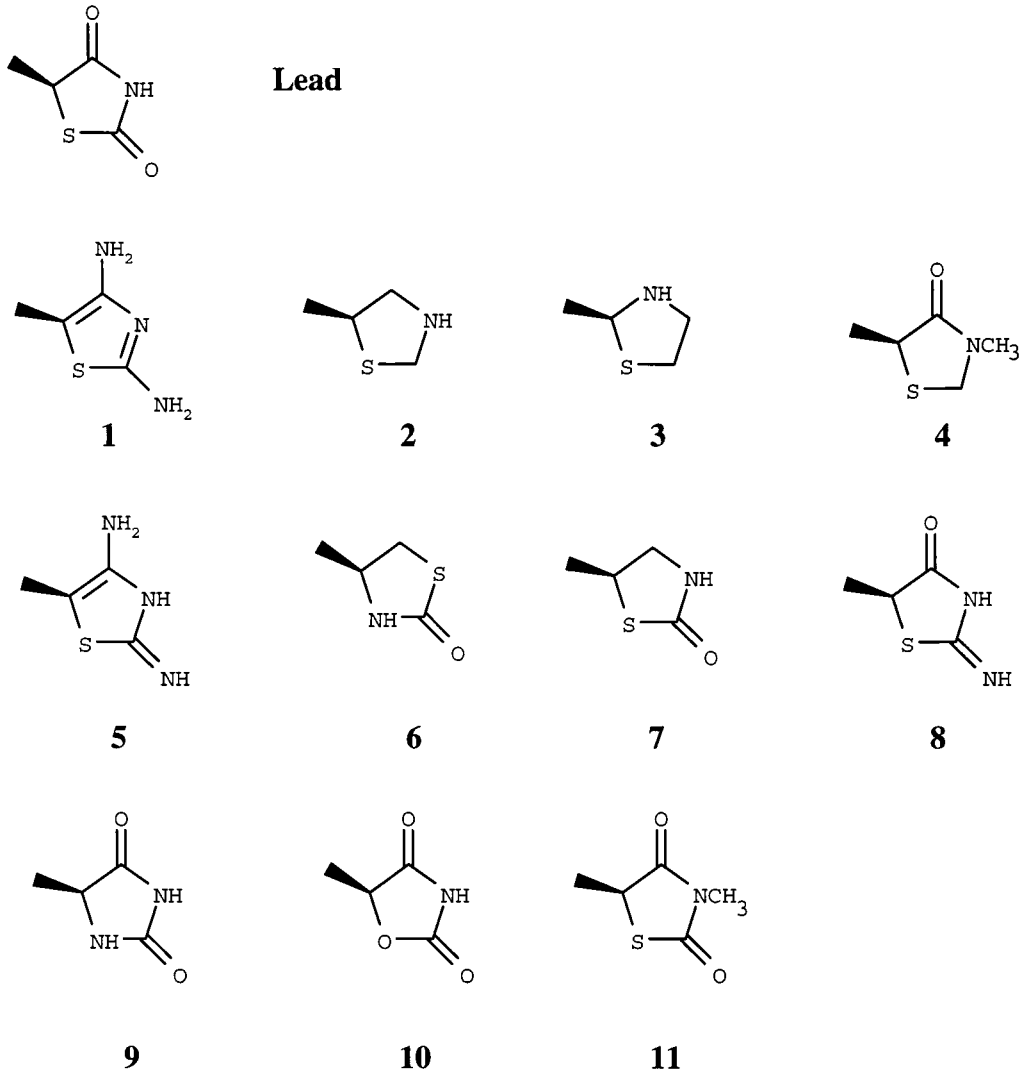
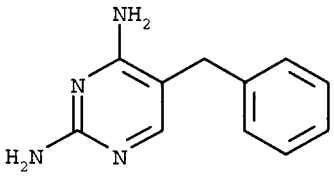


FIGURE 1. Structures of hypoglycemic active lead fragment and ring analogues.

The results demonstrate that the use of the steepest decent procedure for the optimization of molecular similarity does produce, in most cases, the same optimal alignment of the analogues as the simplex method. This is clear from the correlation between the similarity measures obtained from the simplex best calculations and that from the steepest descent procedure. The overall correlation coefficient between these two measures is 0.93 with a root mean square (rms) deviation of 0.062. Only five of the molecules appeared to have located significantly different similarity maxima: hypoglycemic fragment ring analogues 4 and 5, dihydrofolate analogue 16, and serotonin analogues 22 and 23. Removing these five molecules from the analysis gives a correlation coefficient of 0.999 and an rms deviation of 0.009. The fact that different local maxima were located in these five cases is seen from trajectories of nonoptimal simplex runs that appeared to get stuck at the same position in many cases. The location of local stationary points is a well known failing of gradient-based optimization methods and of the relatively primitive steepest descent method in particular. Other gradient-based methods, such as conjugate gradient approaches, would suffer similar drawbacks, although performing a few runs with different randomized starting configurations might easily overcome this problem. However, what Table II clearly illustrates is that for the vast majority of cases



For the lead compound, and 10 analogs, the benzene ring was substituted as follows:

Lead	3,5-$(OCH_3)_2$,4-Br
12	H
13	3-F
14	4-NH_2
15	4-OH
16	4-OCF_3
17	4-$N(CH_3)_2$
18	3-OCH_3
19	3-CF_3
20	3-OCH_3, 4-OH
21	3,5-$(OCH_3)_2$, 4-$(NCH_3)_2$

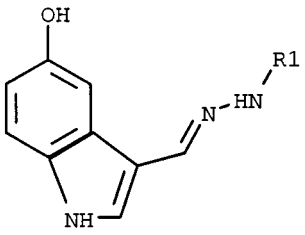
FIGURE 2. Structures of dihydrofolate analogues.

virtually identical similarity maxima are obtained from the best simplex result and that produced by a single steepest descent run.

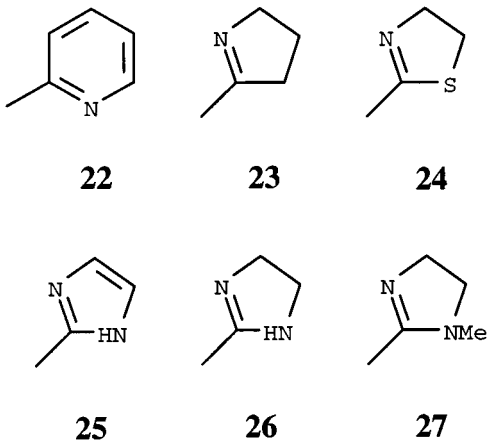
Table III compares the time taken to optimize the similarity using different methods. The simplex average column gives the average time taken to converge for 100 different runs. The simplex best column gives the time taken for the simplex run that produced the best similarity value. Both times are measured relative to the time taken using the gradient-based approach. Typical run times for the gradient-based calculations were less than a second.

The results clearly show the improvement in speed obtained using the steepest descent procedure. Compared to the simplex average results the steepest descent method is approximately 11 times faster, while it is approximately 8 times faster than the best simplex result. Thus we can conclude that the steepest descent method is approximately an order of magnitude faster than the simplex approach. There is only a single exception to this, the dihydrofolate analogue 21. This is due to the gradient method not converging, a failure of the steepest descent algorithm that is likely to be improved using more sophisticated gradient-based approaches.

The time-consuming part of the similarity calculation is the evaluation of the exponential terms of eq. (4). However, the additional calculation of gradients does not involve any further evaluations of exponential functions, and hence there are virtually no extra computational overheads. This, plus the fewer number of similarity evaluations required in any single run, leads to the enhanced performance of the steepest descent optimization.



Where R1 is as follows for the serotonin analogues



22 23 24

25 26 27

FIGURE 3. Structures of serotonin analogues.

Conclusion

A method for optimizing the Carbó and Hodgkin molecular similarity indices using a steepest descent algorithm was described. Its performance for the Carbó index was compared with the simplex optimization procedure. In the vast majority of cases, the new method is considerably faster. Frequently both methods locate the same maximum value.

The more rapid calculation of optimized molecular similarity will enable faster data base searching. As previously discussed,[24] the initial superposition of structures poses a problem for data base searching software. Previous methods require structural features in the data base query to be aligned, resulting in an extra data base search. If our method of optimization was to be incorporated in the search software, then we feel this problem could become redundant. Alternatively the saving in time could be utilized by performing more in-depth calculations on smaller sets of molecules, such as comparing more than one molecular property or similarity index.

TABLE II.
Electrostatic Potential Similarity Values of Compounds Obtained Using Different Optimization Methods.

	Simplex Average	Simplex Best	Steepest Descent
Hypoglycemic ring fragment analogues			
1	0.338	0.379	0.375
2	0.473	0.493	0.493
3	0.405	0.479	0.439
4	0.515	0.635	0.443
5	0.541	0.617	0.487
6	0.636	0.636	0.643
7	0.644	0.659	0.658
8	0.827	0.827	0.827
9	0.858	0.858	0.858
10	0.885	0.884	0.884
11	0.896	0.896	0.896
Dihydrofolate analogues			
12	0.730	0.730	0.730
13	0.648	0.744	0.744
14	0.715	0.715	0.714
15	0.653	0.706	0.706
16	0.555	0.555	0.664
17	0.727	0.731	0.730
18	0.768	0.771	0.771
19	0.485	0.493	0.492
20	0.776	0.780	0.780
21	0.905	0.905	0.905
Serotonin analogues			
22	0.493	0.540	0.369
23	0.586	0.614	0.537
24	0.567	0.597	0.596
25	0.570	0.574	0.563
26	0.544	0.544	0.544
27	0.602	0.602	0.602

Current opinion suggests that shape may be a useful descriptor for the repulsive force between receptor and ligand, with electrostatic potential being the main contributor to the attractive forces. Comparison of both molecular properties may lead to more accurate (in terms of correlation with experimentally determined data) similarity calculations. Our gradient-based optimization methods were implemented into shape similarity calculations using Gaussian functions, and significant performance enhancements were also observed in this area.

Our program that performs this gradient-based optimization is currently command driven, although we are developing a user-friendly X-windows interface. In addition to the implemen-

TABLE III.
Relative Times Taken for Optimization.

	Simplex Average	Simplex Best	Steepest Descent
Hypoglycemic ring fragment analogues			
1	8	4	1
2	7	4	1
3	10	5	1
4	9	3	1
5	7	4	1
6	4	3	1
7	8	3	1
8	6	2	1
9	6	3	1
10	4	3	1
11	9	3	1
Dihydrofolate analogues			
12	14	13	1
13	28	12	1
14	9	6	1
15	34	18	1
16	4	2	1
17	14	12	1
18	6	5	1
19	15	10	1
20	3	3	1
21	1	0.5	1
Serotonin analogues			
22	4	4	1
23	23	25	1
24	10	23	1
25	21	18	1
26	22	16	1
27	13	7	1

tation of gradient-based optimization of shape similarity, it is our intention to incorporate other measures of similarity, as well as improved conformational searching algorithms.

Acknowledgments

A. J. M. is supported by an EPSRC Quota studentship. The work was also funded by a College Research Grant from Birkbeck College.

References

1. W. F. van Gunsteren, P. M. King, and A. E. Mark, *Q. Rev. Biophys.*, **27** 435 (1994).
2. B. Odell, *J. Comput. Aided Mol. Design*, **2**, 191, (1988).
3. A. Seri−Levy, R. Salter, S. West, and W. G. Richards, *Eur. J. Med. Chem.*, **29**, 687 (1994).
4. A. C. Good, S. J. Peterson, and W. G. Richards, *J. Med. Chem.*, **36**, 2929 (1993).
5. A. Seri−Levy, S. West, and W. G. Richards, *J. Med. Chem.*, **37**, 1727 (1994).
6. A. C. Good, S. So, and W. G. Richards, *J. Med. Chem.*, **36**, 433 (1993).
7. M. A. Johnson and G. M. Maggiora, *Concepts and Application of Molecular Similarity*, Wiley, New York, 1990.
8. P. Willet, *Similarity and Clustering in Chemical Information Systems*, Wiley, New York, 1987.
9. P. M. Dean, Ed., *Molecular Similarity in Drug Design*, Chapman Hall, London, 1995.
10. E. E. Hodgkin and W. G. Richards, *Int. J. Quantum Chem.: Quantum Biol. Symp.*, **14**, 105 (1987).
11. C. Burt, W. G. Richards, and P. Huxley, *J. Comput. Chem.*, **11**, 1139 (1990).
12. C. Burt and W. G. Richards, *J. Comput.-Aided. Mol. Design*, **4**, 231 (1990).
13. J. A. Nelder and R. Mead, *Comput. J.*, **7**, 308 (1965).
14. R. Carbó, L. Leyda, and M. Arnau, *Int. J. Quantum Chem.*, **17**, 1185 (1980).
15. A. C. Good, E. E. Hodgkin, and W. G. Richards, *J. Chem. Inf. Comput. Sci.*, **32**, 188 (1992).
16. A. Szabo and N. S. Ostlund, *Modern Quantum Chemistry*, McGraw-Hill, New York, 1989.
17. W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes*, Cambridge University Press, Cambridge, U.K., 1989.
18. A. C. Good and W. G. Richards, *J. Chem. Inf. Comput. Sci.*, **33**, 112 (1993).
19. PIMMS Molecular Modelling System V1.42, Oxford Molecular LTD, Oxford, U.K., 1993.
20. M. J. S. Dewar and W. Thiel, *J. Am. Chem. Soc.*, **99**, 4899 (1977).
21. B. H. Besler, K. M. Merz, Jr., and P. A. Kollman, *J. Comput. Chem.*, **11**, 431 (1990).
22. J. J. P. Stewart, MOPAC 6.0, QPCE 455, Indiana University, 1991.
23. K. H. Buchheit, R. Gamse, R. Giger, D. Hoyer, F. Klein, E. Klöppner, H. Pfannkuche, and H. Mattes, *J. Med. Chem.*, **38**, 2331 (1995).
24. A. C. Good, E. E. Hodgkin, and W. G. Richards, *J. Comput.-Aided. Mol. Design*, **6**, 513 (1992).